Devoir

Exercice 1 : problème du collectionneur (4 pts)

Dans chaque paquet de céréales, une vignette est offerte. Il y en a trois différentes à collectionner.

1. Après avoir obtenu la première vignette, combien de paquets en moyenne doit-on acheter pour obtenir une vignette différente ?
   Indication : on reconnaîtra une loi usuelle.
2. Combien ensuite faudra-t-il acheter de paquets en moyenne pour obtenir la dernière vignette ?
3. Généraliser au cas de n vignettes : quel est le nombre moyen de paquets qu'il faut acheter pour avoir une collection complète de n vignettes ?

On achète des paquets jusqu'à ce qu'on ait obtenu toutes les vignettes : cela ressemble à un problème de loi géométrique. Commençons par bien fixer le cadre. On suppose que les vignettes présentes dans les paquets successifs sont indépendantes et que les 3 vignettes différentes sont équiprobables.

Après avoir obtenu une première vignette, on a donc pour chaque paquet une probabilité de $2/3$ d'obtenir une des deux autres vignettes. Si on note X_2 le nombre de paquets nécessaire pour obtenir une vignette différente, alors X_2 suit la loi géométrique de paramètre $2/3$. Son espérance est $3/2$ et il faut donc acheter en moyenne 1,5 paquets de plus pour obtenir deux vignettes différentes.

Une fois cette seconde vignette obtenue, on a à chaque nouveau paquet une probabilité $1/3$ d'obtenir la troisième vignette. Si on note X_3 le nombre de paquets nécessaire pour cela, alors X_2 suit la loi géométrique de paramètre $1/3$. Son espérance est 3 et il faut en moyenne 3 paquets supplémentaires pour obtenir la collection complète. On obtient ainsi un total moyen de $1 + 3/2 + 3 = 5,5$ paquets.

Généralisons avec n vignettes différentes équiprobables et indépendantes. Une fois k vignettes différentes obtenues, le nombre de paquets nécessaire pour obtenir une $k+1$-ème est la loi géométrique de paramètre $(n-k)/n$ dont l'espérance est $n/(n-k)$. On obtient que le nombre total moyen de paquets à acheter pour obtenir une collection complète est

$$N_{moyen} = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{2} + \frac{n}{1} = n \sum_{j=1}^{n} \frac{1}{j}.$$

Exercice 2 : méthode des moindres carrés (16 pts)

On sait que deux grandeurs x et y sont proportionnelles : $y = ax$ avec $a \in \mathbf{R}$. On cherche à déterminer la valeur de a.

On s'est donné deux valeurs pour x : $x_1 = 2$ et $x_2 = 4$. Puis on a mesuré les valeurs y_1 et y_2 correspondantes. Mais ces mesures sont perturbées par un bruit supposé aléatoire. Leurs valeurs sont ainsi décrites par des variables aléatoires

$$Y_1 = 2a + \varepsilon_1 \quad \text{et} \quad Y_2 = 4a + \varepsilon_2,$$

où ε_1 et ε_2 sont deux variables indépendantes de même densité f définie sur $\mathbf{R}$, d'espérance nulle et de variance 1 : $\mathbf{E}(\varepsilon) = 0$ et $\mathbf{V}(\varepsilon) = 1$.

I. Estimation de a

On cherche à estimer a à partir des valeurs obtenues pour Y_1 et Y_2. Comme a représente le rapport entre y et x, on pourrait logiquement estimer a par les rapports $\frac{Y_1}{x_1}$ et $\frac{Y_2}{x_2}$, ou pourquoi pas, par la moyenne des deux : $\frac{1}{2}(\frac{Y_1}{2} + \frac{Y_2}{4})$. Mais est-ce la meilleure façon de procéder ?

On propose d'estimer a par la valeur

$$A = \alpha Y_1 + \beta Y_2,$$

où α et β sont des nombres réels. On impose les conditions suivantes pour que l'estimation soit la meilleure possible :

$$\mathbf{E}(A) = a \quad \text{et} \quad \mathbf{V}(A) \text{ est minimal.}$$

1. En quoi les conditions imposées impliquent que A permette d'obtenir une bonne estimation de a ?

   On veut estimer a. Le fait que A vaille en moyenne a semble être un minimum. Ce n'est peut-être pas nécessaire mais c'est une condition naturelle. On aimerait surtout que la valeur de A soit la plus proche de a. Autrement dit, on veut que A se disperse le moins possible autour de a. C'est exactement ce que mesure la variance et on désire donc que $\mathbf{V}(A)$ soit minimal.

2. Montrer que les conditions impliquent que $\alpha = \frac{1}{10}$ et $\beta = \frac{1}{5}$.

   $A = \alpha Y_1 + \beta Y_2 = (2\alpha + 4\beta)a + \alpha\varepsilon_1 + \beta\varepsilon_2$. Donc $\mathbf{E}(A) = (2\alpha+4\beta)a + \alpha\mathbf{E}(\varepsilon_1) + \beta\mathbf{E}(\varepsilon_2) = (2\alpha + 4\beta)a$. Pour que $\mathbf{E}(A) = a$ il faut que $2\alpha + 4\beta = 1$.
   D'autre part $\mathbf{V}(A) = \mathbf{V}(\alpha\varepsilon_1) + \mathbf{V}(\beta\varepsilon_2) = \alpha^2\mathbf{V}(\varepsilon_1) + \beta^2\mathbf{V}(\varepsilon_2) = \alpha^2 + \beta^2$.
   On souhaite donc minimiser $\alpha^2 + \beta^2$ avec la contrainte $2\alpha + 4\beta = 1$. Cela revient à minimiser $(\frac{1}{2} - 2\beta)^2 + \beta^2$. Une simple étude (avec une dérivation par exemple) permet d'obtenir $\beta = \frac{1}{5}$ puis $\alpha = \frac{1}{10}$.

3. On obtient $\beta > \alpha$. Cela signifie que pour estimer a, on donne plus de poids à la mesure Y_2 qu'à la mesure Y_1. Expliquer brièvement pourquoi c'est judicieux pour obtenir une meilleure estimation de a.

   Cela peut bien se comprendre graphiquement. On cherche une droite passant par l'origine. Si on se place à une abscisse proche de 0, une incertitude sur l'ordonnée peut beaucoup modifier la pente de la droite correspondante. En revanche à une

abscisse élevée, l'incertitude sur l'ordonnée ne peut modifier que légèrement la pente de la droite. C'est pourquoi la valeur de Y_2 fournit une information plus sûre sur la valeur de a que la valeur Y_1. On lui accorde donc plus de poids.

Remarque : les valeurs obtenues pour α et β correspondent exactement aux valeurs qu'on obtient quand on utilise la méthode des moindres carrés pour obtenir la meilleure droite linéaire à partir des deux points $(2, Y_1)$ et $(4, Y_2)$ obtenus. Notre étude justifie que cette méthode est en un sens la meilleure.

II. Loi de A

On considère désormais la variable $A = \frac{1}{10}Y_1 + \frac{1}{5}Y_2$ obtenue dans la partie précédente.

1. Supposons pour commencer que $\mathcal{E}_1$ et $\mathcal{E}_2$ suivent la loi normale $\mathcal{N}(0,1)$. Quelle est alors la loi de A ?

   $A = a + \frac{1}{10}\varepsilon_1 + \frac{1}{5}\varepsilon_2$. On sait qu'une combinaison linéaire de lois normales indépendantes est encore une loi normale : $\varepsilon_1/10 \sim \mathcal{N}(0, 1/100)$ et $\varepsilon_2/5 \sim \mathcal{N}(0, 1/25)$. Donc $A \sim \mathcal{N}(a, 1/20)$.

2. Déterminer dans ce cas la valeur de η pour laquelle $\mathbf{P}(|A - a| \leqslant \eta) \approx 0,99$.

   On se ramène à la loi $\mathcal{N}(0,1)$: $\frac{A-a}{1/\sqrt{(20)}} \sim \mathcal{N}(0,1)$. Alors $\mathbf{P}(|A-a| \leqslant \eta) = \mathbf{P}(\sqrt{20}|A-a| \leqslant \sqrt{20}\eta) = 2F_{\mathcal{N}}(\sqrt{20}\eta) - 1 \approx 0.99$. Donc $F_{\mathcal{N}}(\sqrt{20}\eta) \approx 0.995$ et $\sqrt{20}\eta \approx 2.58$. On obtient $\eta \approx 0.58$. Autrement dit, on garantit avec notre méthode une précision sur a de l'ordre de 0,6.

3. Dans le cas général, on détermine la loi de A en utilisant la densité f de $\mathcal{E}_1$ et $\mathcal{E}_2$. Soit $t \in \mathbf{R}$. Déterminer l'expression de la fonction de répartition $F_A(t)$ sous forme intégrale en fonction de la densité f.

   $A = a + \frac{1}{10}\varepsilon_1 + \frac{1}{5}\varepsilon_2$. Alors $\mathbf{P}(A \leqslant t) = \mathbf{P}(\frac{1}{10}\varepsilon_1 + \frac{1}{5}\varepsilon_2 \leqslant t - a)$.
   Les variables ε_1 et ε_2 sont indépendantes de même densité f. L'évènement ci-dessus correspond graphiquement à la région du plan située sous la droite d'équation $\frac{x}{10} + \frac{y}{5} = t - a$. Ainsi

$$\mathbf{P}(A \leqslant t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{5(t-a)-x/2} f(x)f(y)dydx.$$

IV. Généralisation

On effectue maintenant n mesures en n points $x_1, \ldots, x_n$. On obtient à chaque fois une valeur $Y_i = ax_i + \varepsilon_i$. La méthode des moindre carrés permet d'estimer a par la variable

$$A = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

On aimerait déterminer le nombre de points n qu'il faut considérer pour assurer avec un risque de 1% que $|\tilde{A} - a| < 10^{-2}$.

1. Donner $\mathbf{E}(A)$ et $\mathbf{V}(A)$.

$\mathbf{E}(A) = \frac{\sum_{i=1}^n x_i \mathbf{E}(Y_i)}{\sum_{i=1}^n x_i^2}$. Or $\mathbf{E}(Y_i) = ax_i + \mathbf{E}(\mathcal{E}_i) = ax_i$, donc $\mathbf{E}(A) = \frac{\sum_{i=1}^n x_i^2 a}{\sum_{i=1}^n x_i^2} = a$.
De même, en utilisant $\mathbf{V}(Y_i) = \mathbf{V}(\mathcal{E}_i) = 1$ et le fait que les variables $x_i Y_i$ sont indépendantes, $\mathbf{V}(A) = \frac{\sum_{i=1}^n \mathbf{V}(x_i Y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sum_{i=1}^n x_i^2 \mathbf{V}(Y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{1}{\sum_{i=1}^n x_i^2}$.

2. Pour quelle raison les variables $x_i Y_i$ ne satisfont pas les hypothèses du théorème central limite ?

   $x_1 Y_1 = ax_1^2 + x_1 \mathcal{E}_1^2$ et $x_2 Y_2 = ax_2^2 + x_2 \mathcal{E}_1^2$ ne suivent pas la même loi si $x_1 \neq \pm x_2$ (on voit par exemple que leurs espérances ax_1^2 et ax_2^2 sont différentes). Les variables $x_i Y_i$ bien qu'indépendantes ne sont donc pas iid et le théorème ne s'applique pas.

3. À l'aide de l'inégalité de Bienaymé-Tchebychev, donner une condition sur les x_i qui garantisse la précision demandée.

   On souhaite obtenir $\mathbf{P}(|A - a| \leqslant 10^{-2}) > 0,99$. Prenons l'évènement contraire : $\mathbf{P}(|A - a| > 10^{-2}) \leqslant 0,01$. Or d'après l'inégalité de Bienaymé-Tchebychev, $\mathbf{P}(|A - a| > 10^{-2}) \leqslant \frac{\mathbf{V}(A)}{(10^{-2})^2}$. Si $\mathbf{V}(A) \leqslant 0,01 \cdot 10^{-4} = 10^{-6}$ on obtient le résultat attendu. Or $\mathbf{V}(A) = \frac{1}{\sum_{i=1}^n x_i^2}$ et on peut donc conclure que si $\sum_{i=1}^n x_i^2 \geqslant 10^6$, on est certain d'obtenir une approximation de a à 10^{-2} près avec un risque de 1%.

   Afin de pouvoir utiliser le TCL, nous allons simplifier notre problème en considérant tous les nombres x_i égaux à x. La variable A devient alors $A = a + \frac{1}{nx} \sum_{i=1}^n \varepsilon_i$.

4. À l'aide du théorème central limite, déterminer une condition sur n et x qui permette d'obtenir la précision demandée.
   Qu'obtient-on pour $x = 2$ et $x = 4$?

   Les variables $\mathcal{E}_i$ sont iid d'espérance 0 et variance 1. Si n est assez grand, on peut considérer que $\frac{\sum_{i=1}^n \varepsilon_i - 0}{1 \cdot \sqrt{n}}$ suit à peu près la loi $\mathcal{N}(0,1)$.
   Alors $\mathbf{P}(|A - a| \leqslant 10^{-2}) = \mathbf{P}(|\frac{1}{nx} \sum_{i=1}^n \varepsilon_i| \leqslant 10^{-2}) = \mathbf{P}(|\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i| \leqslant 10^{-2} x\sqrt{n}) \approx 2F_{\mathcal{N}}(10^{-2} x\sqrt{n}) - 1$. On souhaite obtenir une probabilité de l'ordre de $0,99$. On obtient $10^{-2} x\sqrt{n} \approx 2,58$ et donc $n \approx \frac{6,66 \cdot 10^4}{x^2}$.
   Pour $x = 2$, on obtient $n \approx 16600$ et pour $x = 4$, $n \approx 4200$. Moins de valeurs sont nécessaires pour $x = 4$ car on a vu que plus x est grand, plus la valeur de Y obtenue fournit une estimation fiable de a.